同济大学

毕业设计(论文)开题报告

（适用于工科类、理科类专业）

课题名称	整合核小体定位信息预测非典型 DNA 位点
副 标 题	
学院（系）	生命科学与技术学院
专　　业	生物信息专业
学生姓名	张佳胜　学　号　1753096

2020 年 月 日

一、毕业设计（论文）课题背景（含文献综述）

典型性 DNA (B-form DNA, B-DNA)，指的是右手双螺旋结构的 DNA。随着科研的进步，DNA 的二级结构不断被发现，出现了一部分特殊的 DNA 结构，研究者们称之为非典型 DNA (non-B-form)[2]。非典型 DNA 包括 Z-DNA、十字形 DNA (cruciforms)、三链体 DNA (triplexes) 以及四链体 DNA (G4)。由于非典型 DNA 的特定作用，使得它也作为基因组结构研究的一个重要方面，被研究者们广泛关注。

四链 DNA 结构，被称为 G4s (G-quadruplexes)[1][3]，又叫鸟嘌呤四链体，一般产生于 G 四平方面的自我堆叠。四平方面通常由四个鸟嘌呤产生的环 Hoogsten 氢键排列而形成的，在四平方面的中心，存在一个一价阳离子稳定结构。

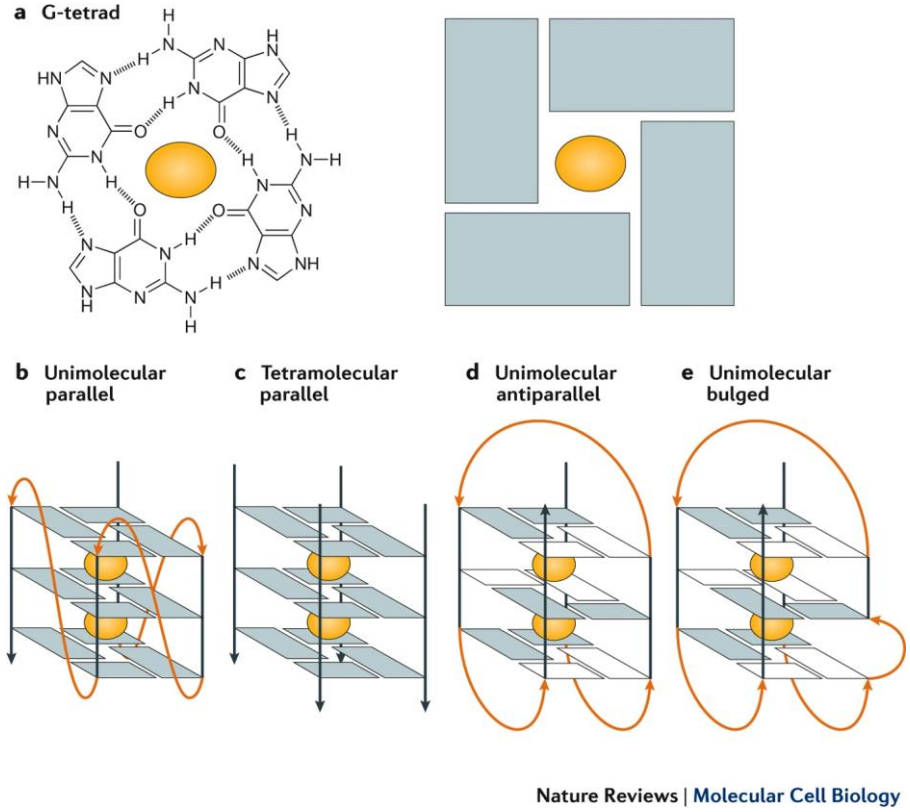

Figure 1 | G4s 结构[1]。由四个鸟嘌呤通过环 Hoogsten 氢键排列而形成四平方面，四平方面的中心由金属阳离子稳定结构，然后通过链的不同排列方式将四平方面堆叠，构成所谓的 G4s。

G4 最早发现于体外，后来也被证实普遍存在于体内[2]。研究发现，基因组上 G4 结构位点并不随机，而且具有特定序列模式[7] (科学家们总结出的典型的 G4 保守模式: $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$)，常出现在部分关键性基因区域，尤其是基因的启动子区域和端粒区域。进一步的研究表明，G4s 可能在基因表达调控以及端粒作用等方面扮演着重要的角色[4]。本课题主要针对 G4s 结构位点，探究整合核小体定位信息辅助预测 G4s 结构位点的算法。

过去的一些实验技术手段[5][7]，如核磁共振、X 射线等方法，只能够针对少量的序列和片段证实其形成 G4s 结构的能力，无法在基因组层面大通量地预测新的 G4s 结构位点。为了更好地研究 G4s，研究者们发展 G4s 的测序技术，尝试在基因组中对 G4s 结构进行定位。

现在采用的比较完整的 G4s 的测序技术，是通过染色体免疫共沉淀 (G4 ChIP-seq) 的方式来探测有 G4s 结构的序列，并由此产生了一部分 G4 ChIP-seq 的数据。为了大通量地获取全基因组的 G4s 结构位点，研究者们希望使用计算机预测的方式针对这部分 G4 ChIP-seq 的数据进行分析。他们根据 G4s 的特定序列模式以及在序列上的分布特点，开发出了一些算法和工具。现有的主要算法和工具可以分为三大类[7]：通过 G4 序列模式进行正则匹配、滑动窗口结合打分矩阵进行评估预测以及机器学习的方式。前两者[8-10]基于实验经验，根据之前实验研究获取的 G4 的特定序列模式，特殊结构位点等信息来进行模型的构建；而后者[13]则完全由数据驱动，基于大量的标记好的序列训练预测模型，达到最终的预测效果。但是，这些算法各自因为匹配模式的宽泛程度在不同细胞间存在差异以及不同细胞系的数据的数量和质量有所差异等种种原因，很难针对不同的细胞中的 G4s 结构位点进行准确预测。所以，我们希望能够找到一种算法模型，可以解析不同组织细胞中 G4s 的结构位点图谱及其异质性，并且提升 G4s 结构位点预测的准确性。

核小体是由 DNA 和组蛋白形成的染色质的基本结构单位。由于核小体定位的相关性研究比较多，科学家们已经开发出了一套比较完整的实验操作(MNase-seq)及其数据分析流程。MNase-seq[16]，采用限制性外切酶将染色质上不受保护的区域 (开放区域) 切除，留下核小体上缠绕的 DNA 序列，然后针对这部分序列进行扩增和测序。我们能够从公共数据库内获取大量的核小体的 MNase-seq 的数据，并且根据一些数据分析的流程筛选高质量的数据，然后我们可以通过 MNase-seq 的测序结果回帖到基因组上，找出可能的核小体定位，获取核小体的定位信息。

研究者发现，核小体定位与 G4s 位点存在一定的相关性，而且可以借助一些拟合算法构建核小体定位与 G4s 位点的关联，这表明核小体定位的信息可能可以辅助序列模式来预测 G4s 位点，改进现有的一些计算方法。由于实验室开发出的核小体 MNase-seq 数据分析以及质量检测流程的算法 CAM[14]的存在，我们能够借助这一工具以及已有的一部分高质量的整合核小体定位数据辅助 G4s 位点的预测，保证了我们的预测数据的质量。同时，我们能够拿到大量的不同物种不同细胞类型的比较完备的全基因组核小体定位数据，根据这部分数据结合已知的算法来构建 G4s 结构位点的预测模型，帮助我们解析不同细胞类型中的 G4s 结构位点图谱，预测它的位点。这很好地弥补了我们之前提到的 G4 ChIP-seq 数据在不同细胞中不足以及质量较差的问题。

ATAC-seq[16]通过 Tn5 转座酶随机插入染色质开放区域，将不受核小体保护的染色质区域(染色质开放区域) 切割下来，通常用于定位染色质开放区域。部分研究认为[15]，当 Tn5 转座酶切割的两端正好跨过一个核小体的时候，可以认为切割下来的这个长片段中能够定位到核小体。那么通过筛选 ATAC-seq 数据的序列长度 (>146bp)，mapping 后获取这些序列的覆盖信息，就能够从另一个角度获取核小体定位。基于现如今 ATAC-seq 技术的发展以及 ATAC-seq 数据的普及性，我们希望能够同时借助 ATAC-seq 的数据来分析核小体定位，提升我们的数据质量。故而，我们将对 ATAC-seq 和 MNase-seq 的核小体占位信息进行综合性比较，考虑是否采取 ATAC-seq 数据来进一步扩大和完善我们的核小体定位信息。

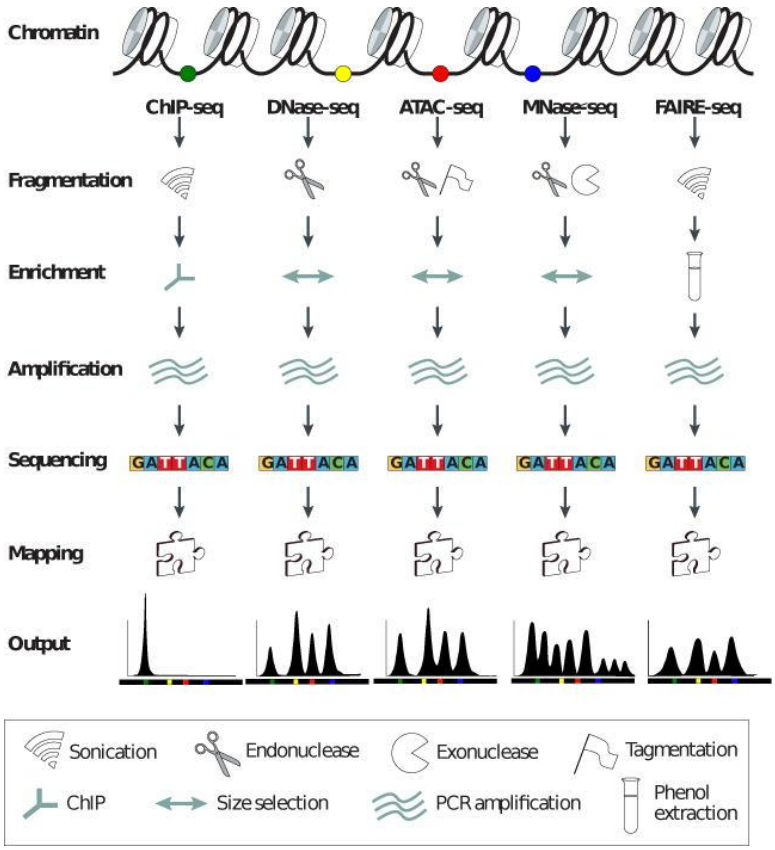

Figure 2 | 二代测序技术[16]。二代测序技术的基本示意图，其中包括了我们之前提到过的 ChIP-seq、ATAC-seq 以及 MNase-seq 技术。

综上所述，本次课题将从以下几个方面展开：首先，我们获取 MNase-seq 和 ATAC-seq 的数据，基于各自的分析流程获取核小体占位信息 (nucleosome occupancy)，进行综合性分析比较；其次，我们将分析核小体定位和 G4s 结构位点之间的关系，证明其相关性并构建算法，用于辅助预测 G4s 结构位点；通过我们的整合核小体数据辅助现有的一些序列模式匹配的算法来构建新的预测模型，实现异质性细胞的 G4s 结构位点图谱的绘制以及针对不同细胞的位点预测；然后我们会整理现有的不同算法，以及它们对 G4s 结构位点的预测效果，指出我们的模型优势。最后，针对模型，我们进行必要的验证以及可靠性评估。

二、毕业设计（论文）方案介绍（主要内容）

本项目的主体部分是构建从核小体定位数据辅助预测 G4s 位点的算法模型，主要分为三个阶段：数据准备阶段，算法模型构建阶段，模型预测效果评估以及验证阶段。我们将按照以下步骤展开内容：

(一)MNase-seq 数据和 ATAC-seq 数据推断的核小体占位的系统性比较

首先从 NCBI 上获取 MNase-seq 和 ATAC-seq 的数据，采用 wget 的方法下载对应细胞系的 SRR 文件，并使用 faster-dump 转换成 fastq 文件。

然后，针对 MNase-seq 的 fastq 文件，我们采用已经构建好的数据分析流程，通过安装好的 CAM 程序进行核小体定位的综合分析，获取对应的质检报告以及分析结果。针对

ATAC-seq 的数据，我们先进行 fastQC 质量检测，针对质量较高的测序文件进行下一步分析，剔除接头后，我们使用 bowtie 将这些序列 mapping 到基因组上，获取相应的 bed 和 bam 文件，之后将 bam 文件连同参考基因组输入到 nucleoATAC[15]，运行脚本，获取分析结果。

针对两套数据(事实上是四组数据，人类 K562 各一组，小鼠神经前体细胞各一组)的核小体定位 profile 文件进行核小体占位分析综合性的比较。

(二)MNase-seq 和 G4 ChIP-seq 数据的获取，清洗和整合

同理，利用 wget 方法从 NCBI 上获取 G4 ChIP-seq 数据。然后采用传统的测序数据分析流程，进行质检，清洗以及序列回帖。根据物种以及细胞系整理好不同的数据。

(三)整合核小体定位与 G4s 位点的相关性探究

针对同一种细胞系比较 MNase-seq 获取的核小体定位数据与 G4 ChIP-seq 获取的 G4s 结构位点，综合进行分析，绘制对应的 XY 相关性图、热图以及占位图。

(四)整合核小体定位数据预测 G4s 位点的算法调研、比较和选取

调研部分预测算法，如回归拟合(包括线性回归以及逻辑回归)、决策树、支持向量机以及贝叶斯模型等，选取不同的预测算法拟合核小体定位数据和 G4s 位点的关系，比较其准确率、敏感率等学习指标，选取最符合期望的预测算法作为我们模型构建的算法。

(五)利用整合核小体数据辅助现有 motif 算法构建预测模型

调研现有的 motif 算法以及其他 G4 结构位点预测算法，如 G4Hunter[8]、G4Killer[9]以及 pqsfinder[10]等等，包括这些算法文献的阅读、算法使用以及预测效果。

利用之前构建的核小体定位数据预测 G4s 位点的算法，进行模型的初步筛选，然后利用现有的一些算法构建预测模型的第二层筛选。完成模型后，采用不同细胞系构建训练集和测试集获取预测指标，完善模型以及模型参数。

(六)比较不同模型的预测效果

比较我们构建的核小体定位数据辅助预测模型和已有的模型的预测效果，针对不同细胞系的 G4s 位点进行预测，比较预测的准确率、ROC 曲线等等指标。总结并归纳各个算法的优劣性以及预测表现，然后撰写报告。

(七)绘制细胞异质性 G4s 结构位点图谱

根据构建好的预测模型，对不同生物的不同细胞系的细胞进行 G4s 的位点预测，获取不同细胞系的 G4s 结构位点。比较并整合这些 G4s 结构位点数据，绘制细胞异质性 G4s 结构位点图谱。

（八）验证模型的准确性与可靠性

使用新的数据集针对模型进行准确性的检验，检验预测的假阳性率、假阴性率、准确度以及敏感度等指标。评估我们模型预测不同细胞系 G4s 结构位点的可靠性并撰写报告。

三、毕业设计（论文）的主要参考文献

[1] Hänsel-Hertsch R, Di Antonio M, Balasubramanian S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential [J]. Nat Rev Mol Cell Biol. 2017 May; 18(5):279-284.

[2] Bochman ML, Paeschke K, Zakian VA. DNA secondary structures: stability and function of G-quadruplex structures [J]. Nat Rev Genet. 2012 Nov; 13(11):770-80.

[3] Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology [J]. Nucleic Acids Res. 2015 Oct 15; 43(18):8627-37.

[4] Mukherjee AK, Sharma S, Chowdhury S. Non-duplex G-Quadruplex Structures Emerge as Mediators of Epigenetic Modifications [J]. Trends Genet. 2019 Feb; 35(2):129-144.

[5] Chambers VS, Marsico G, Boutell JM, et al. High-throughput sequencing of DNA G-quadruplex structures in the human genome [J]. Nat Biotechnol. 2015 Aug; 33(8):877-81.

[6] Hänsel-Hertsch R, Beraldi D, Lensing SV, et al. G-quadruplex structures mark human regulatory chromatin [J]. Nat Genet. 2016 Oct; 48(10):1267-72.

[7] Puig Lombardi E, Londoño-Vallejo A. A guide to computational methods for G-quadruplex prediction [J]. Nucleic Acids Res. 2020 Jan 10; 48(1):1-15.

[8] Brázda V, Kolomazník J, Lýsek J, et al. G4Hunter web application: a web server for G-quadruplex prediction [J]. Bioinformatics. 2019 Sep 15; 35(18):3493-3495.

[9] Brazda V, Kolomaznik J, Mergny JL, et al. G4Killer web application: a tool to design G-quadruplex mutations [J]. Bioinformatics. 2020 May 1; 36(10):3246-3247.

[10] Hon J, Martínek T, Zendulka J, et al. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R [J]. Bioinformatics. 2017 Nov 1; 33(21):3373-3379.

[11] Yang D. G-Quadruplex DNA and RNA [J]. Methods Mol Biol. 2019; 2035:1-24.

[12] Kwok CK, Merrick CJ. G-Quadruplexes: Prediction, Characterization, and Biological Application [J]. Trends Biotechnol. 2017 Oct; 35(10):997-1013.

[13] Garant JM, Perreault JP, Scott MS. G4RNA screener web server: User focused interface for RNA G-quadruplex prediction [J]. Biochimie. 2018 Aug; 151:115-118.

[14] Hu S, Chen X, Zhang Y, et al. CAM: A quality control pipeline for MNase-seq data. PLoS One. 2017 Aug 7; 12(8):e0182771.

[15] Alicia N. Schep, 1 Jason D. Buenrostro, 1 Sarah K. Denny, et al. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res. 2015 Aug 27; 25: 1757-1770.

[16] Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet. 2014 Nov;15(11):709-21.

四、审核意见

指导教师审核意见：（针对选题的价值及可行性作出具体评价） 课题《整合核小体定位信息预测非典型 DNA 位点》采用整合的核小体定位信息构建预测 G4s 结构位点的算法，能够从基因组层面大通量的预测 G4s 结构位点，解决实验检测 G4s 的部分难题。目前已具备比较完备的整合核小体定位信息，能够支持进行多细胞系的细胞异质性 G4s 结构位点图谱的绘制。课题各阶段的时间安排也较为合理，故该课题也有着较高可行性。综上，我同意张佳胜以此课题开题。 指导教师签名_____ _____年_____月 _____日
专业审核意见： 负责人签名_____ _____年_____月_____日